An Overview of Microarray Data Analysis at NIEHS

Microarray Users' Group

April 21, 2000

Components of Analysis

- Data acquisition and image analysis
- Statistical analysis of ratio data to identify "outlier" genes
- Follow-up analysis (clustering, etc.) on the set of differentially expressed genes

Analysis Software

IPLabs (Scanalytics, Inc.)

Image processing

ArraySuite (Yidong Chen/Scanalytics)

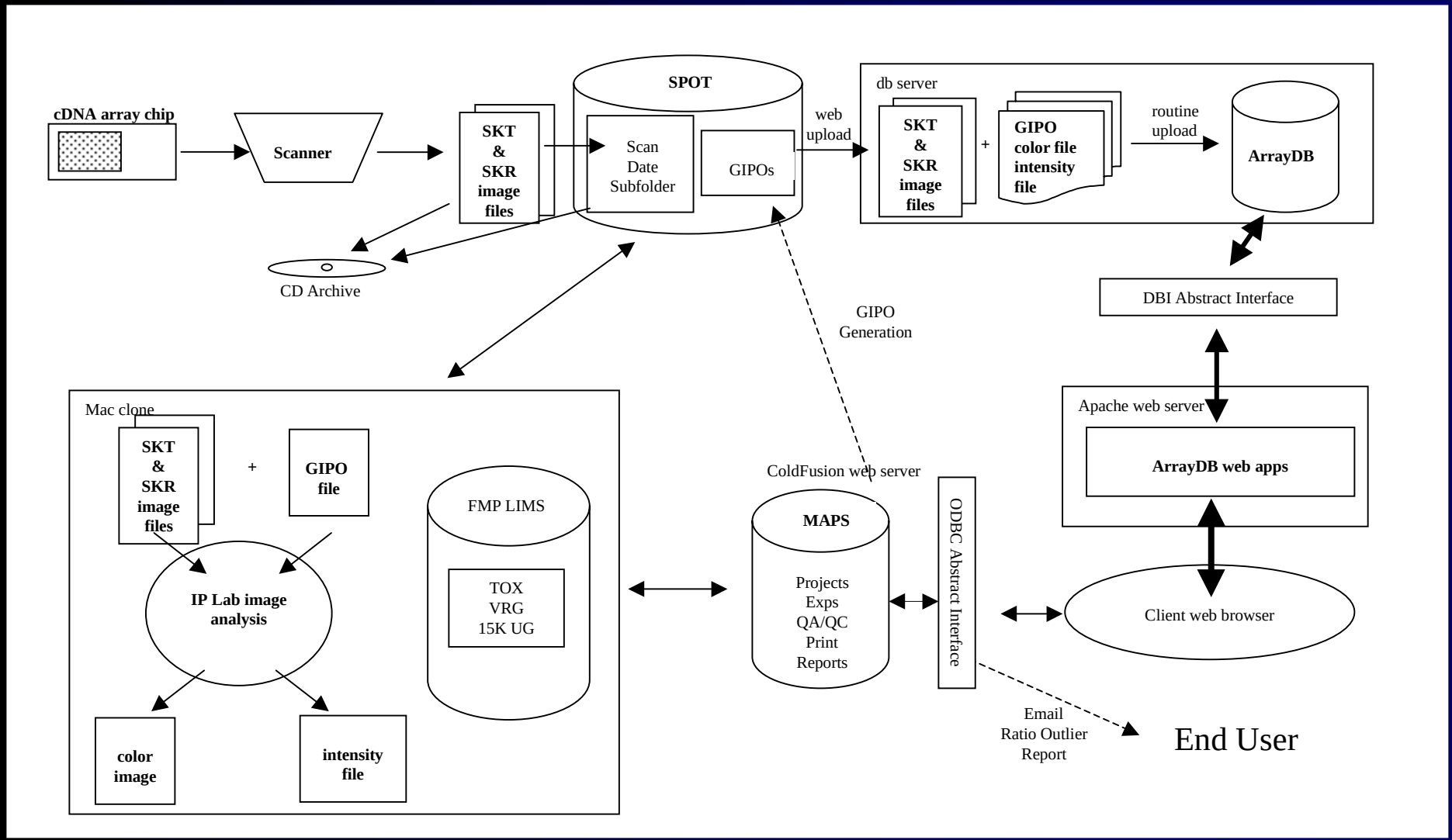
Array alignment, target location, & ratio analysis

Cluster (Stanford University)

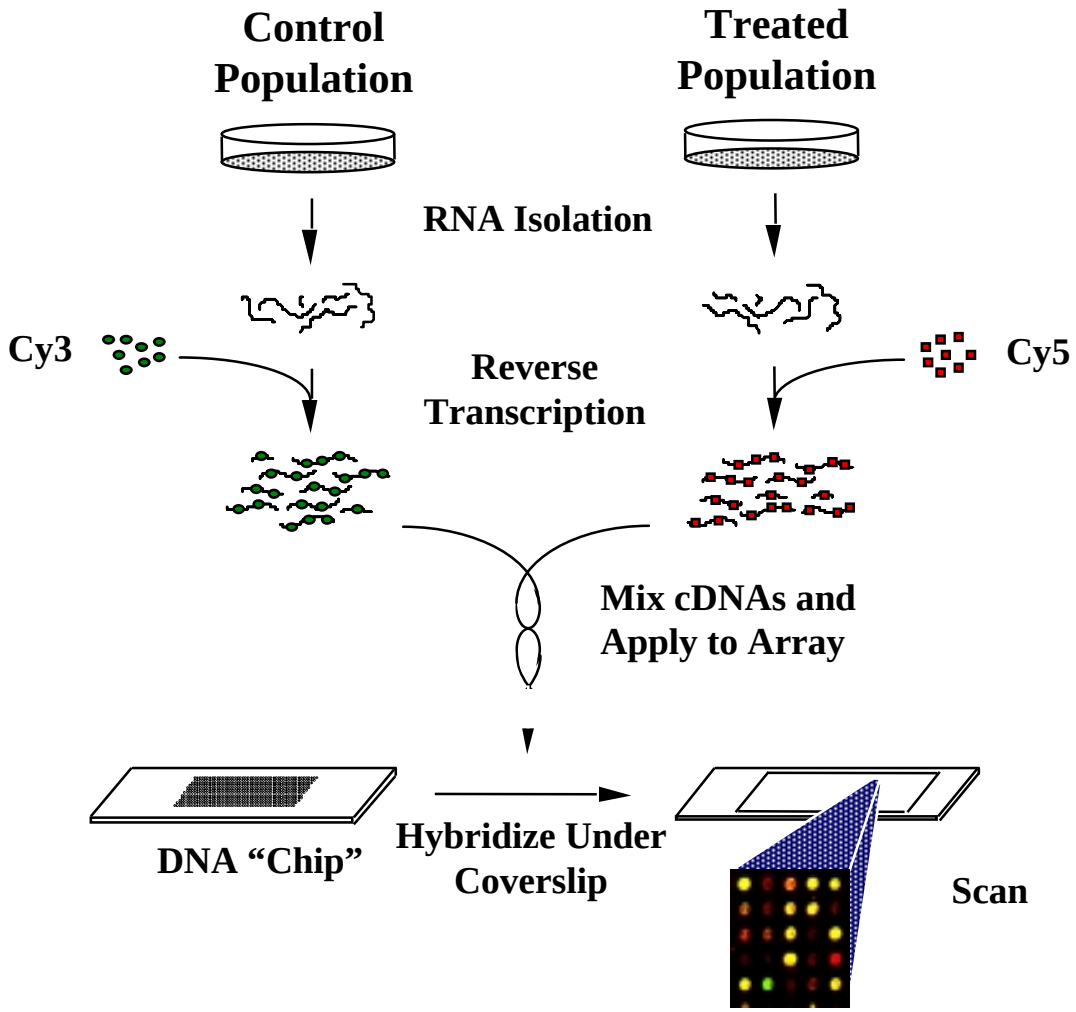
Spotfire (Spotfire AB)

GeneSpring (Silicon Genetics)

NMC Data Management



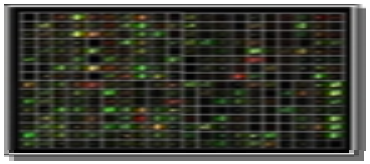
Simplified Overview of Gene Expression Analysis Using cDNA Microarrays



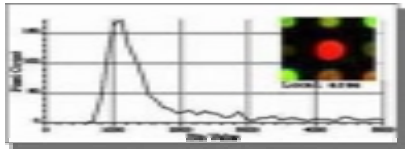
Sources of Variability

- Within-array variation
  - Labeling and hybridization differences
- Between-array variation
  - Print to print differences
- Biological variability
  - RNA quality
  - Expression differences between animals
- Measurement error (scanning)

Image Analysis

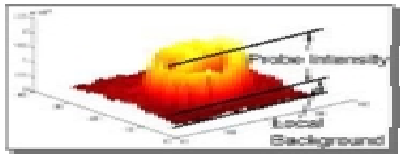


1. Target Segmentation

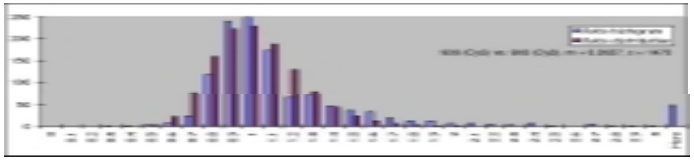


2. Background Subtraction

3. Target Detection



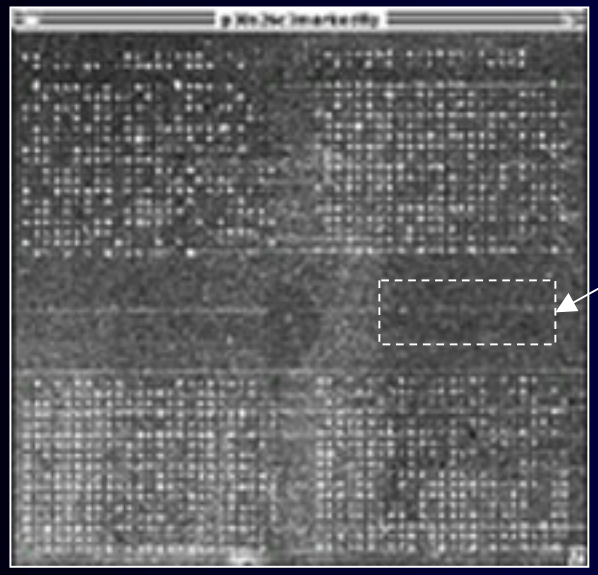
4. Target Intensity Determination



5. Ratio Analysis

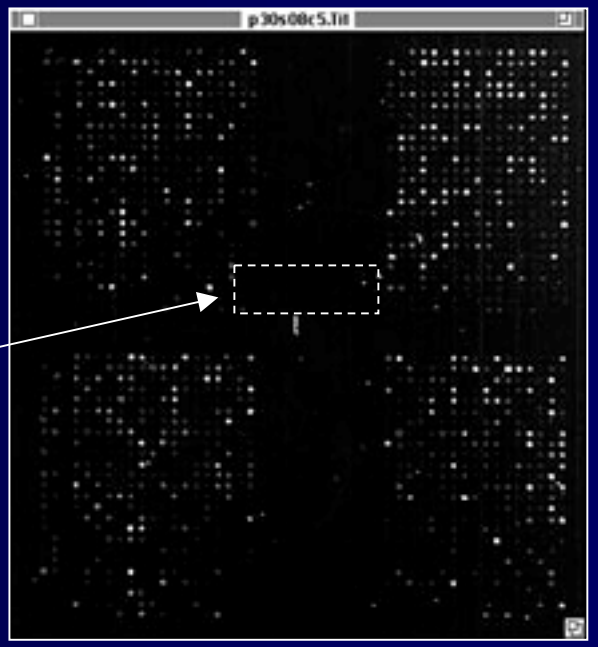
Images courtesy of Dr. Y. Chen, NHGRI

Background Variation

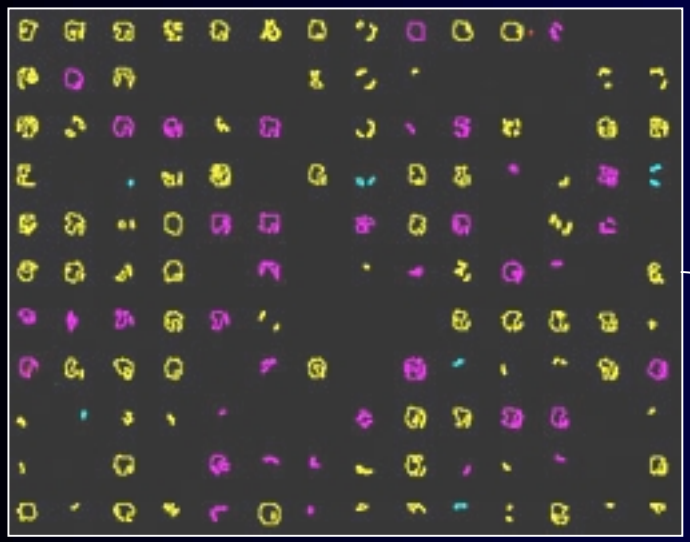


Background Mean: 345
Std. Dev.: 158

Background Mean: 187
Std. Dev.: 250

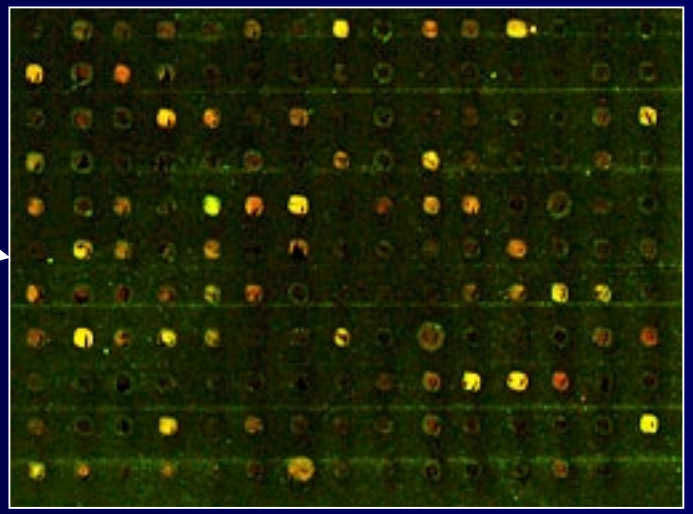


Locating Array Targets



Pixels significantly more intense than the background are selected to define the target area

Pseudo-color image is generated according to intensity ratio

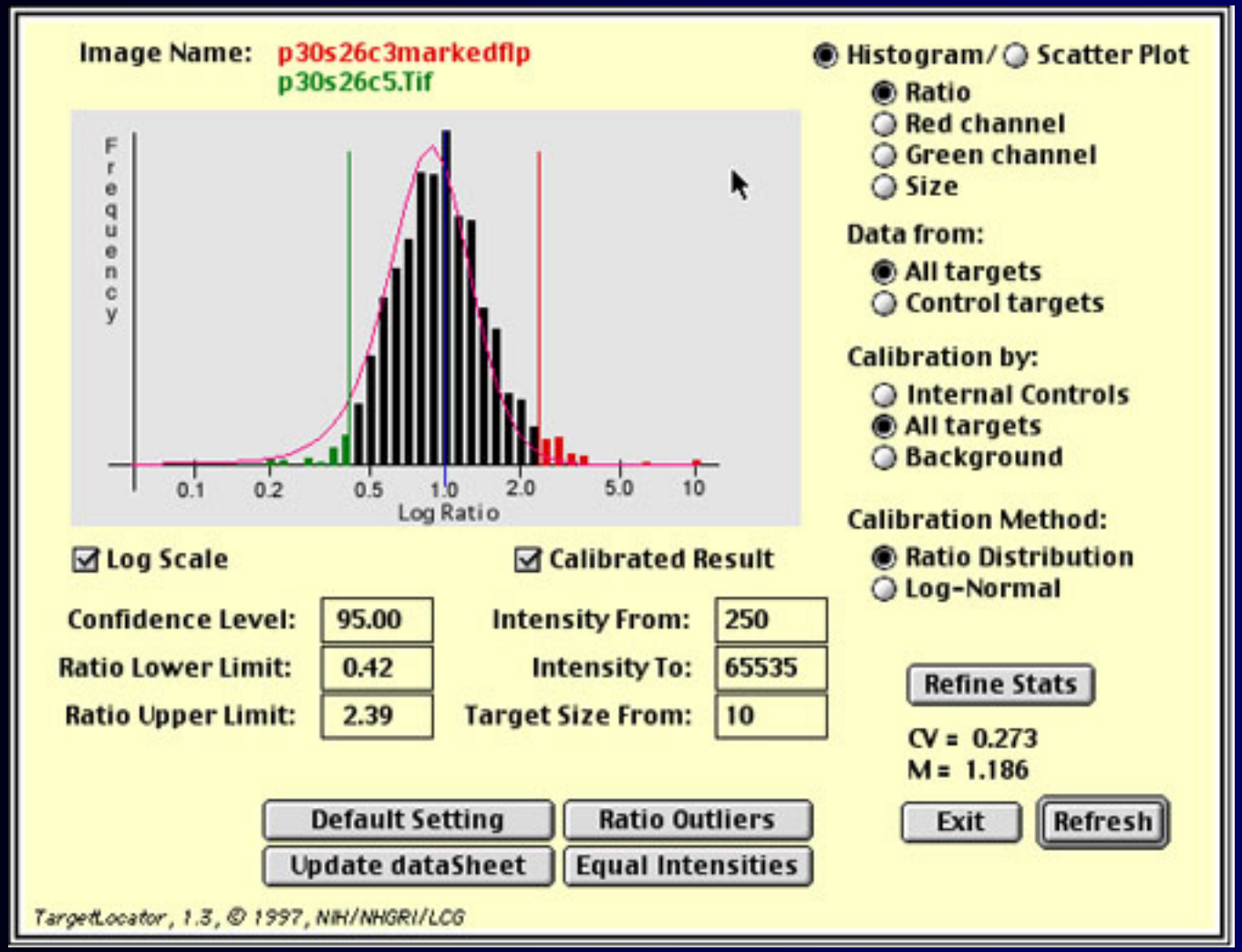


Statistical Analysis

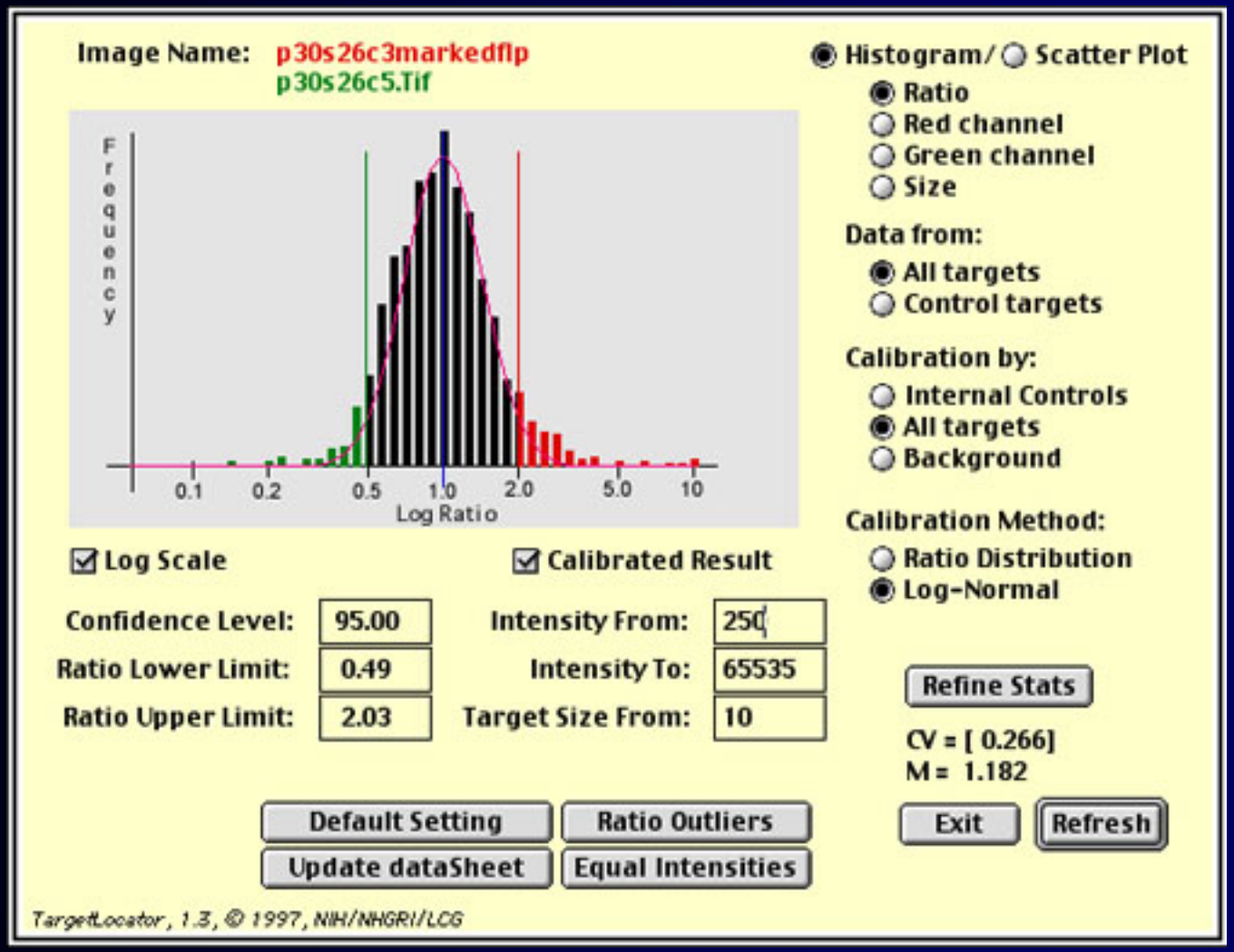
- Intensity ratios have a constant *coefficient of variation*[1].
  - Variability in a response increases as the response itself increases
- Analyze the *logarithm* of the ratios.
  - Evens out skew distributions
  - Gives values that are more independent of the absolute magnitude of the response

[1]Chen *et al.* J. Biomedical Optics **2**(4), 1997.

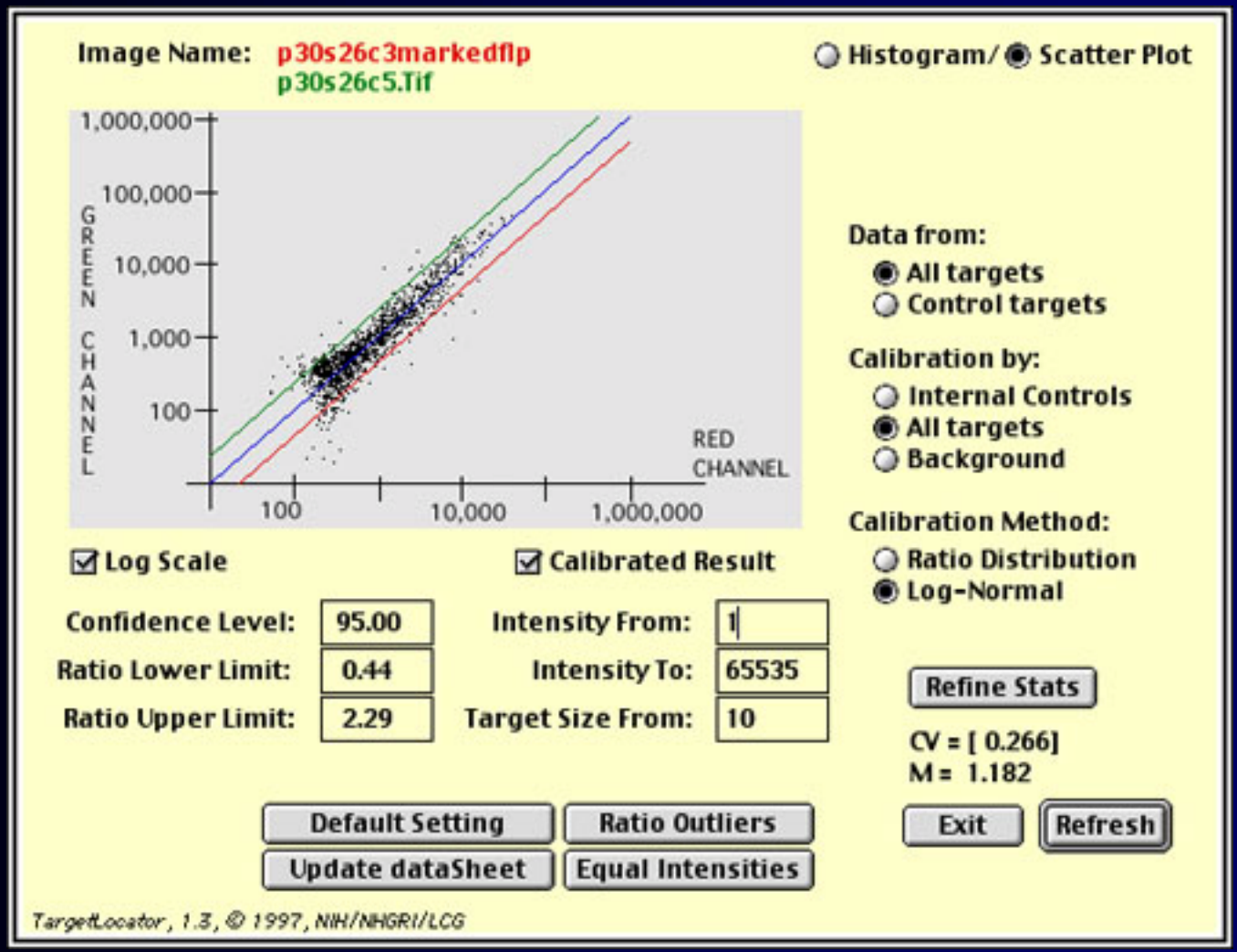
Analysis - Ratio Distribution



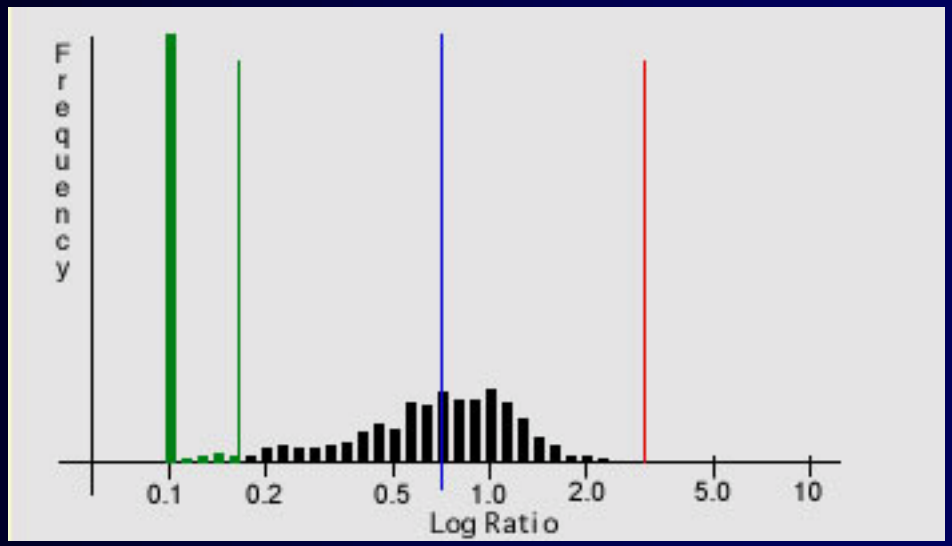
Analysis - Lognormal Distribution



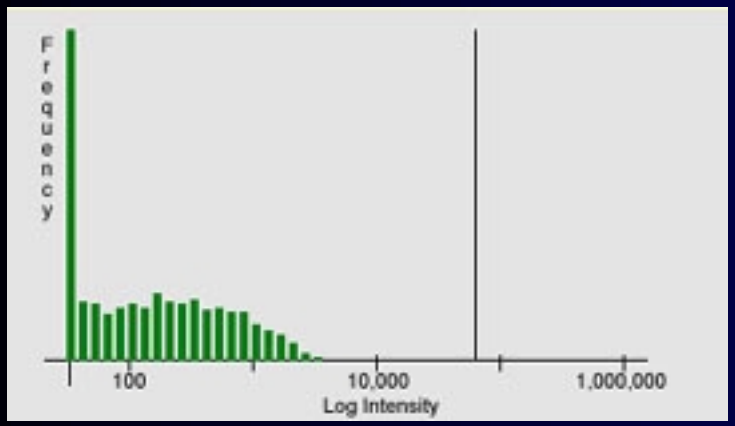
Scatterplot of ratio data



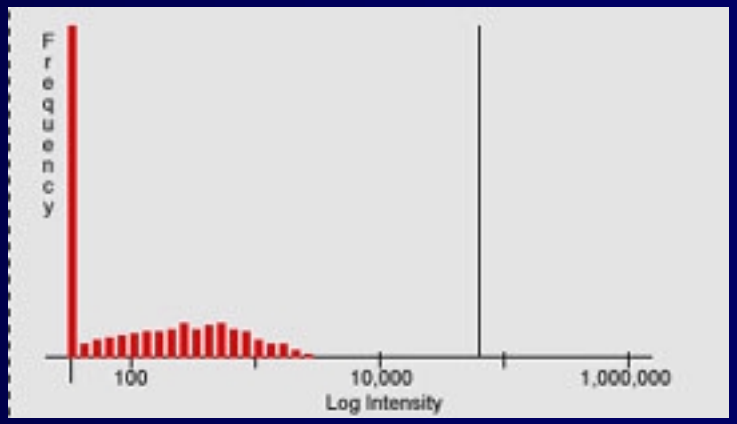
Another example



Green channel

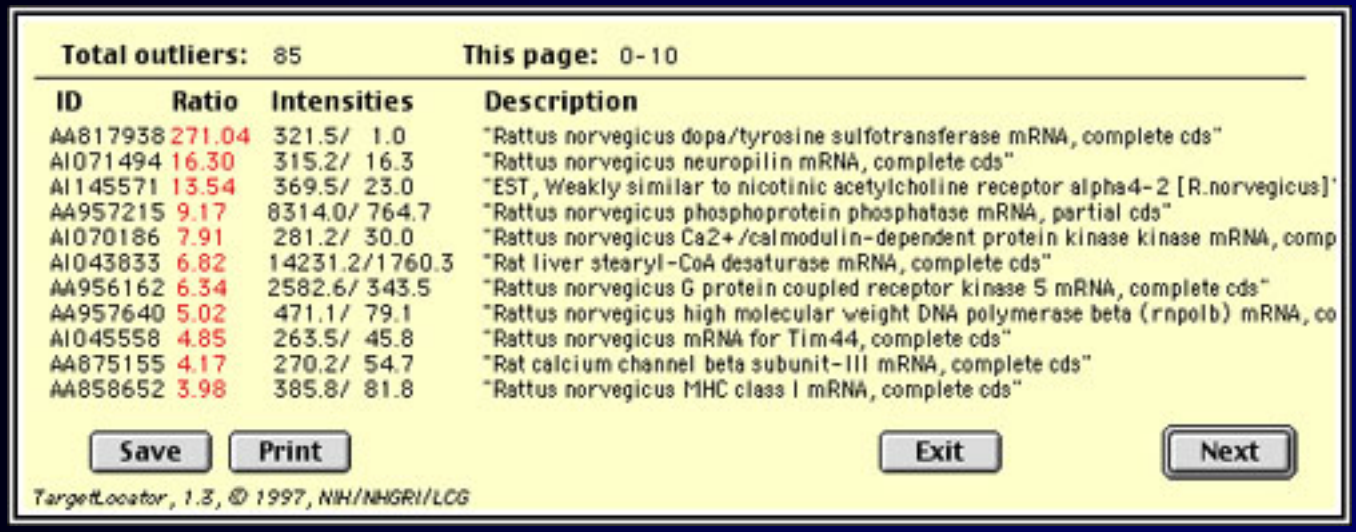


Red channel



Identifying Outlier Genes

For a specified *confidence level* (95%, 99%, etc.)
a list of "outlier" genes is produced.



Total outliers: 85 **This page:** 0-10

ID	Ratio	Intensities	Description
AA817938	271.04	321.5/ 1.0	"Rattus norvegicus dopa/tyrosine sulfotransferase mRNA, complete cds"
AI071494	16.30	315.2/ 16.3	"Rattus norvegicus neuropilin mRNA, complete cds"
AI145571	13.54	369.5/ 23.0	"EST, Weakly similar to nicotinic acetylcholine receptor alpha4-2 [R.norvegicus]"
AA957215	9.17	8314.0/ 764.7	"Rattus norvegicus phosphoprotein phosphatase mRNA, partial cds"
AI070186	7.91	281.2/ 30.0	"Rattus norvegicus Ca2+/calmodulin-dependent protein kinase kinase mRNA, comp
AI043833	6.82	14231.2/1760.3	"Rat liver stearyl-CoA desaturase mRNA, complete cds"
AA956162	6.34	2582.6/ 343.5	"Rattus norvegicus G protein coupled receptor kinase 5 mRNA, complete cds"
AA957640	5.02	471.1/ 79.1	"Rattus norvegicus high molecular weight DNA polymerase beta (rnpolb) mRNA, co
AI045558	4.85	263.5/ 45.8	"Rattus norvegicus mRNA for Tim44, complete cds"
AA875155	4.17	270.2/ 54.7	"Rat calcium channel beta subunit-III mRNA, complete cds"
AA858652	3.98	385.8/ 81.8	"Rattus norvegicus MHC class I mRNA, complete cds"

Save Print Exit Next

TargetLocator, 1.3, © 1997, NIH/NHGRI/LCG

Replication

- Why replicate?
  - Genes may be identified as differentially expressed in an experiment *completely at random*.
  - The chances of this depend on the *confidence level* used in the analysis.
- How many times?
  - Depends primarily on the availability of resources
  - Also depends on what level of uncertainty in your results you can tolerate

Replication

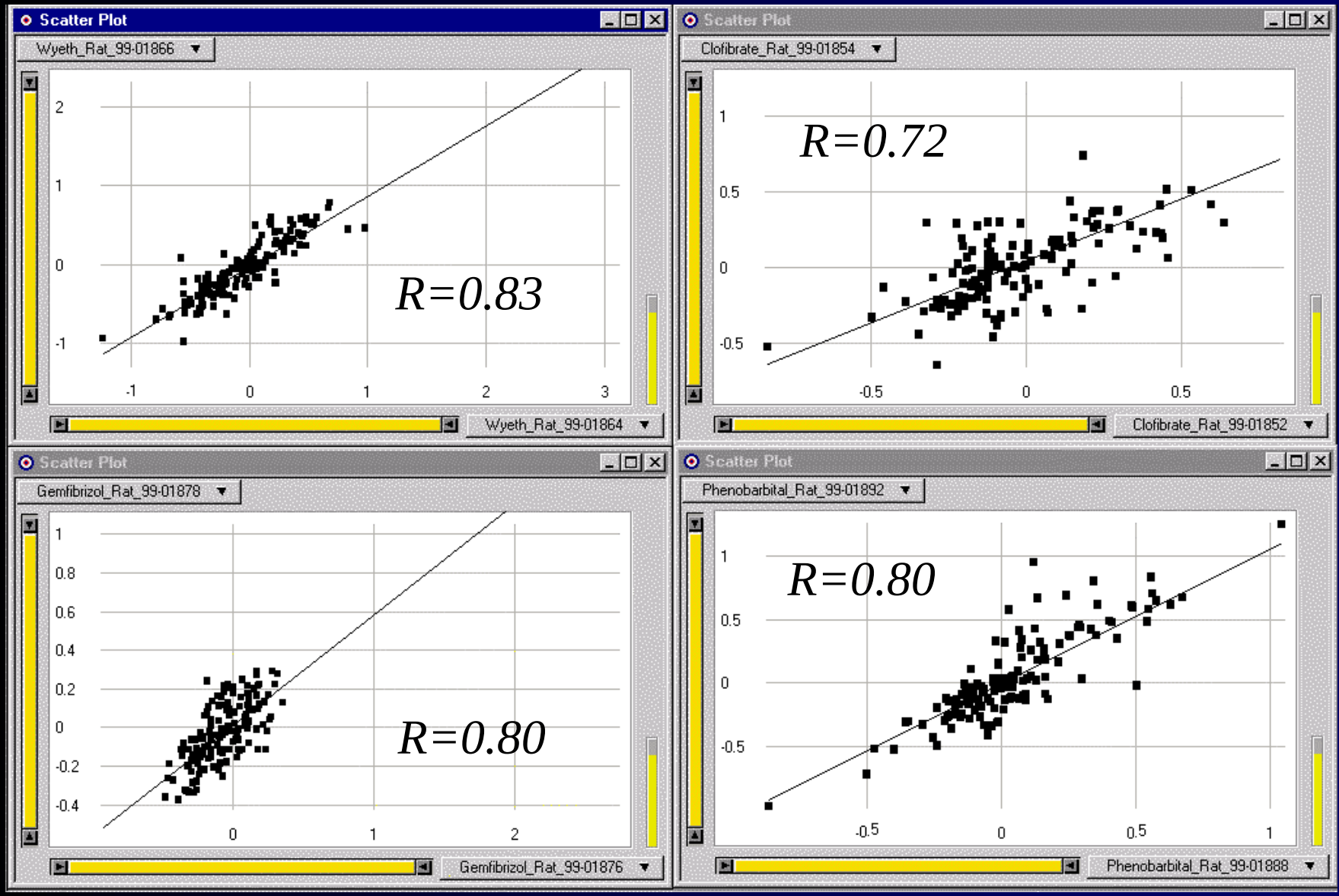
Three replicates at 95% confidence, 2000 gene array

Times flagged by chance	*Probability (p)*	*Frequency (Np)*
0	0.8574	1715
1	0.1354	271
2	0.007125	14
3	0.000125	<1

Three replicates at 95% confidence, 12000 gene array

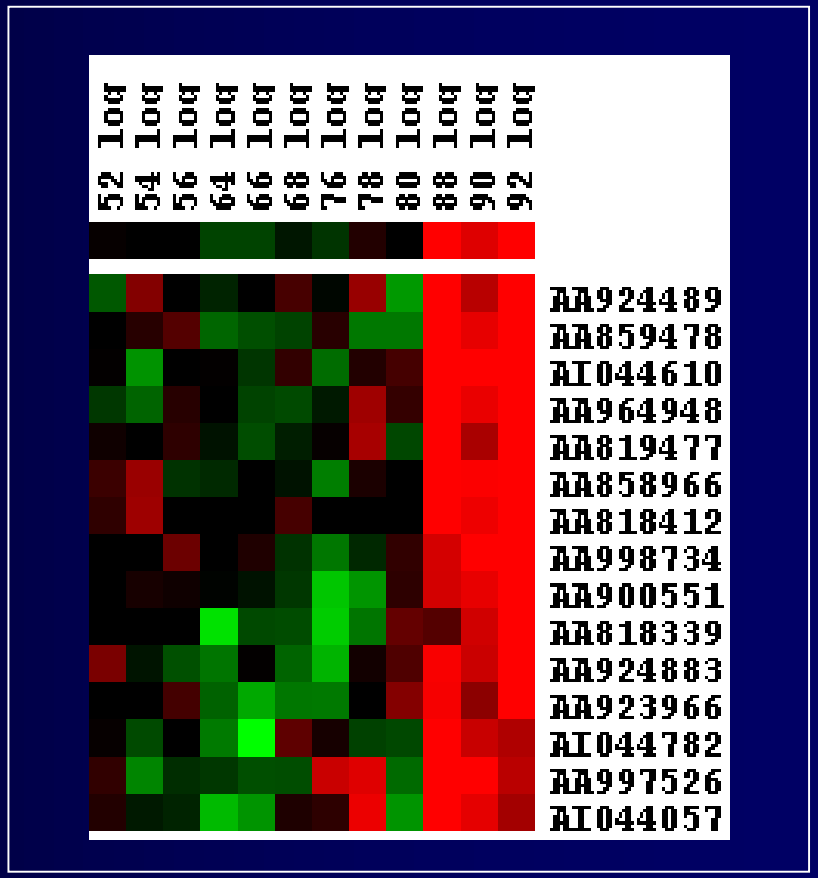
Times flagged by chance	*Probability (p)*	*Frequency (Np)*
0	0.8574	10289
1	0.1354	1625
2	0.007125	85
3	0.000125	1

Correlations between animals



Cluster Analysis

- Allows identification of groups of genes that are similarly expressed
- Several methods:
  - Hierarchical (trees)
  - Self-organizing maps
  - Gene shaving
  - Support vector machines



Post-Analysis

- Images are stored
  - ArrayDB (under development)
  - Archived to CD
- All processed data will be stored in the ArrayDB database (when the system is fully implemented) for subsequent analysis.